Noname manuscript No.
(will be inserted by the editor)

Is a comparison of results meaningful from the inexact replications of computational experiments? (ELECTRONIC SUPPLEMENTARY MATERIAL)

Matej Črepinšek · Shih-Hsi Liu · Luka Mernik · Marjan Mernik

the date of receipt and acceptance should be inserted later

1 Statistical comparison

While performing statistical tests [7] the first question is which statistical test, parametric or non-parametric, is suitable. It is well known that parametric tests have much more power (probability that a statistical test will correctly reject a false null hypothesis, or a probability of avoiding a Type II error) than a non-parametric test. Hence, we should use parametric tests whenever allowed. Conditions for parametric statistical tests are independence, normality, and homoscedasticity [7]. Since we do not have access to raw data of the experiments performed by Waghmare in [8] we can only check normality (e.g., by Kolmogorov-Smirnov test [7]) of our own raw data and make a

Matej Črepinšek
University of Maribor, Maribor, Slovenia
E-mail: matej.crepinsek@um.si

Shih-Hsi Liu
California State University, Fresno, USA
E-mail: shliu@CSUFresno.edu

Luka Mernik
California Institute of Technology, Pasadena, USA
E-mail: lmernik@caltech.edu

Marjan Mernik
University of Maribor, Maribor, Slovenia
E-mail: marjan.mernik@um.si

conjecture that the same holds for the data in [8]. If assumptions about approximate normal distribution do not hold then we need to resort to less powerful non-parametric tests and working with ordinal/rank-order data instead of interval or ratio data. However, using non-parametric statistics with a low number of algorithms (k) and problems (N) might also be very risky. For example, the Friedman test is appropriate when $K > 5$ and $N > 10$ [3]. For these reasons, some researchers advocate the usage of parametric test with some adjustments when assumptions for parametric test are violated (e.g., claiming bigger Type I Error) [7]: "*The reluctance among some sources to transform interval/ratio data into an ordinal/rank-order or categorical/nominal format for the purpose of analyzing it with a nonparametric test, is based on the fact that interval/ratio data contain more information than either of the latter two forms of data. Because of their reluctance to sacrifice information, these sources take the position that even when there is reason to believe that one or more of the assumptions of a parametric test has been violated, it is still more prudent to employ the appropriate parametric test. ... Under such conditions, however, most researchers would probably conduct a more conservative t test in order to avoid inflating the likelihood of committing a Type I error (i.e., one might employ the tabled critical* $t_{.01}$ *value to represent the* $t_{.05}$ *value instead of the actual value listed for* $t_{.05}$ *).*" In our case, as a reader will notice, we are dealing with a small number of problems N as well, while on the other hand assumptions for parametric tests do not always hold. Due to the aforementioned reasons we will apply both, parametric and non-parametric, tests. It is shown in these particular cases that the results from both tests are mostly the same and that selection between parametric and non-parametric statistical tests was not crucial. The same conclusions were achieved also in [7].

Since the results for mean and standard deviations were presented in [2] and [8] the parametric z-Test [1] is suitable for showing statistical significance when data are normally distributed. The null hypothesis states that the mean values are equal ($H_0 : \mu_1 = \mu_2$), whilst an alternative hypothesis states that the mean values are not equal ($H_1 : \mu_1 \neq \mu_2$). The z-value is calculated by Eq. 2.

$$z = \frac{\overline{Y_1} - \overline{Y_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{2}$$

where $\overline{Y_i}$ represents the i-th sample mean, σ_i the standard deviation of the i-th sample, and n_i the number of independent runs of the i-th sample. From the z-value the p-value is computed by Eq. 3.

$$p = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\Pi}} e^{\frac{-u^2}{2}} du \tag{3}$$

When data are not normally distributed we can apply the Wilcoxon signed-rank non-parametric test [9], which is a common method for comparing two algorithms over multiple data sets. The Wilcoxon test ranks the differences in performances of two algorithms for each problem by absolute values and then compares the ranks between the positive and the negative differences. The difference of the two algorithms on i-th problem is denoted by d_i. The R^+ is the sum of ranks for problems on which the second algorithm outperformed the first (where detected difference in performances is positive), the R^- is the sum of ranks for problems on which the first algorithm outperformed the second (where the detected difference in performances is negative). The ranks of differences that equal 0 are split evenly among the sums (Eq. 4).

$$\begin{aligned} R^+ &= \sum_{d_i>0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i) \\ R^- &= \sum_{d_i<0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i) \end{aligned} \tag{4}$$

$$z = (T - N(N+1)/4)/\sqrt{N(N+1)(2N+1)/24} \tag{5}$$

The statistics z (Eq. 5), where T is the smaller of the sums, $T = \min(R^+, R^-)$, is normally distributed approximately.

1.1 Statistical analysis for constrained problems

Since, the raw data from Tables 1 and 2 mostly failed the Kolmogorov-Smirnov test [7] for normality of distribution (only in the case of $TLBO_{s1}$ of Table 2) we increased the level of significance from 95% to 99% ($\alpha = 0.01$) making the aforementioned adjustments. Note that functions f_1 and f_4 are constrained problems, and feasible solutions are irregularly spread and normality of distribution regarding solutions can

be easily violated. The null hypothesis H_0 cannot be rejected if the p-value is bigger than the significance level α. From Tables 5 and 6 (last column) we can see that the p-value is always bigger than the significance level and null hypothesis H_0 cannot be rejected with, now smaller and adjusted, 95% confidence (not 99% confidence).

Table 5 Parametric z-Test for f_1 (Table 1) [8] (global optimum=-15).

	Črepinšek[2]		Waghmare[8]		Statistics	
Method	Mean	SD	Mean	SD	z-value	p-value
$TLBO_{s1}$	-13.845	±2.4	-13.045	±2.59	-1.24094	> 0.01
$TLBO_{s2}$	-13.864	±1.7	-13.451	±2.47	-0.75441	> 0.01
$TLBO_{s3}$	-13.199	±1.5	-13.633	±1.47	1.13184	> 0.01
$TLBO_{s4}$	-13.743	±1.9	-14.246	±2.26	0.93314	> 0.01

Table 6 Parametric z-Test for f_4 (Table 4)[8] (global optimum=7049.248021).

	Črepinšek[2]		Waghmare[8]		Statistics	
Method	Mean	SD	Mean	SD	z-value	p-value
$TLBO_{s1}$	7257.002	±102.6	7249.525	±142.93	0.232547	> 0.01
$TLBO_{s2}$	7244.565	±95.8	7235.805	±108.99	0.330653	> 0.01
$TLBO_{s3}$	7386.036	±209.2	7348.48	±69.87	0.93274	> 0.01
$TLBO_{s4}$	7252.795	±100	7223.492	±110.757	1.07583	> 0.01

Since, the raw data for Tables 1 and 2 have mostly failed the Kolmogorov-Smirnov test [7] for normality of distribution, some might argue that the aforementioned adjustments are not appropriate. Hence, we have applied also the Wilcoxon signed-rank non-parametric test [9], which is a widely used method (e.g., [5,6]) for the comparison of two algorithms over multiple data sets and we also applied it on f_1 and f_4 (Tables 7 and 8). Note that the number of problems is probably too small ($N = 4$) for reliable non-parametric tests. From both non-parametric Wilcoxon's tests we could not reject the null hypothesis, and the results between [2] and [8] are insignificantly different ($p - value = 1$ for f_1 and $p - value = 0.0679$ for f_4).

Table 7 Non-parametric Wilcoxon test for f_1 (Table 1) [8] (global optimum=-15).

	Čre.[2]	Wag.[8]	Wilcoxon statistics			
Method	Mean	Mean	d	$\|d\|$	rank	
$TLBO_{s1}$	-13.845	-13.045	-0.800	0.800	4	R+ = 5
$TLBO_{s2}$	-13.864	-13.451	-0.413	0.413	1	R- = 5
$TLBO_{s3}$	-13.199	-13.633	0.434	0.434	2	T = 5
$TLBO_{s4}$	-13.743	-14.246	0.503	0.503	3	z = 0

Table 8 Non-parametric Wilcoxon test for f_4 (Table 4)[8] (global optimum=7049.248021).

	Čre.[2]	Wag.[8]	Wilcoxon statistics			
Method	Mean	Mean	d	$\|d\|$	rank	
$TLBO_{s1}$	7257.002	7249.525	7.477	7.477	1	R+ = 10
$TLBO_{s2}$	7244.565	7235.805	8.760	8.760	2	R- = 0
$TLBO_{s3}$	7386.036	7348.48	37.556	37.556	4	T = 0
$TLBO_{s4}$	7252.795	7223.492	29.303	29.303	3	z = -1.825741858

1.2 Statistical analysis for unconstrained problems

We have compared the mean values on the number of fitness evaluations and their standard deviations with the parametric z-Test [1] after checking normal distribution of data, which holds for all functions in Table 3 except for De Jong and Hyper Sphere, where assumptions were slightly violated. Since this parametric test is quite robust it still provides reliable information in spite of the fact that test assumptions have been violated [7]. The results of z-Test are presented in Table 9, where it can be seen that the results are statistically insignificant ($p > 0.01$) for the following functions: Martin and Gaddy, Rosenbrock (D=2) - b, Hyper Sphere (D=6), and Branin. While the z-Test shows statistically significant results for the following functions: De Jong, Goldstein and Price, Rosenbrock (D=2) - a, and Rosenbrock (D=3). Among them it can be noticed that Waghmare's results are significantly better for De Jong, Goldstein and Price, and Rosenbrock (D=3), while it is significantly worse for Rosenbrock (D=2) - a. Since not all data were normally distributed, the aforementioned adjustment and statistical tests can be accepted at 95% confidence level.

Due to the fact that the raw data for Table 3 in few cases did not passed Kolmogorov-Smirnov test [7] for normality of distribution some might argue that the aforementioned adjustments are not appropriate. Hence, we have also applied Wilcoxon's non-

parametric tests for the optimization problems under discussion (Table 10). From this non-parametric test, again we cannot reject that the null hypothesis and results between [2] and [8] are insignificantly different ($p - value = 0.0929$).

Table 9 Parametric z-Test for different optimization methods (Table 8)[8].

	Črepinšek[2]		Waghmare[8]		Statistics	
Function	Mean #FE	SD	Mean #FE	SD	z-value	p-value
De Jong	832.2	±400.1	472	±341.5	3.75058	≤ 0.01
Goldstein and Price	629	±119.7	543.4	±113.7	2.83991	≤ 0.01
Martin and Gaddy	317	±97	280.8	±71.2	1.64781	> 0.01
Rosenbrock (D=2) - a	694.3	±374	954.4	±399.1	-2.60466	≤ 0.01
Rosenbrock (D=2) - b	1911	±884.2	1896	±527.3	0.0798047	> 0.01
Rosenbrock (D=3)	9992.6	±3791.3	4838.4	±1097.8	7.15238	≤ 0.01
Hyper Sphere (D=6)	750.8	±60.2	746.4	±67.4	0.266678	> 0.01
Branin	577.5	±220.1	572	±205.7	0.0999964	> 0.01

Table 10 Non-parametric Wilcoxon test for different optimization methods (Table 8)[8].

	Čre.[2]	Wag.[8]	Wilcoxon statistics			
Function	Mean #FE	Mean #FE	d	$\|d\|$	rank	
De Jong	832.2	472	360.2	360.2	7	R+ = 30
Goldstein and Price	629	543.4	85.6	85.6	5	R- = 6
Martin and Gaddy	317	280.8	36.2	36.2	4	T = 6
Rosenbrock (D=2) - a	694.3	954.4	-260.1	260.1	6	z = -1.680336101
Rosenbrock (D=2) - b	1911	1896	15.0	15.0	3	
Rosenbrock (D=3)	9992.6	4838.4	5154.2	5154.2	8	
Hyper Sphere (D=6)	750.8	746.4	4.4	4.4	1	
Branin	577.5	572	5.5	5.5	2	

We performed a statistical z-Test [1] on those functions having the more different results (e.g., Rosenbrock and Ackley) after checking normal distribution of data, which do not hold for functions in Table 4 except for Rosenbrock (D=30 and D=50). From Table 11 it can be seen that the results are mostly statistically insignificant ($p > 0.01$). Since not all data were normally distributed we did the aforementioned adjustment, and the statistical tests could be accepted at a 95% confidence level. For those which

are statistically significant (Ackley, D=10, 30) there is still a question of the practical importance of such significance (e.g., $4.23E-15$ vs. $3.55E-15$). An interesting discussion between statistical and practical significance can be found in [4].

Since, the raw data for Table 4 have mostly not passed the Kolmogorov-Smirnov test [7] for normality of distribution, some might argue that the aforementioned adjustments are not appropriate. Hence, we have also applied Wilcoxon's non-parametric tests for optimization problems on Rosenbrock and Ackley functions for different dimensions (Table 12). From these two non-parametric tests again we cannot reject the null hypotheses, and the results between [2] and [8] are again insignificantly different ($p-value = 0.2249$ for Rosenbrock and $p-value = p = 0.3452$ for Ackley).

Table 11 Parametric z-Test for different optimization methods (Table 10)[8].

		Črepinšek[2]		Waghmare[8]		Statistics	
Function	D	Mean	SD	Mean	SD	z-value	p-value
Rosenbrock	5	3.55E-02	±3.00E-01	1.76E-03	±2.05E-03	0.615991	> 0.01
Rosenbrock	10	9.38E-02	±1.54E-01	6.92E-02	±1.12E-01	0.70759	> 0.01
Rosenbrock	30	2.16E+01	±9.23E-01	2.17E+01	±1.07E+00	-0.387606	> 0.01
Rosenbrock	50	4.33E+01	±7.43E-01	4.30E+01	±8.10E-01	1.49493	> 0.01
Rosenbrock	100	9.47E+01	±1.04E+00	9.46E+01	±9.63E-01	0.386433	> 0.01
Ackley	5	1.81E-15	±1.57E-15	9.47E-16	±1.60E-15	2.10867	> 0.01
Ackley	10	4.23E-15	±8.48E-16	3.55E-15	±8.02E-31	4.39211	≤ 0.01
Ackley	30	4.48E-15	±3.55E-16	3.55E-15	±8.02E-31	14.3488	≤ 0.01
Ackley	50	2.75E-01	±1.95E+00	4.73E-05	±2.59E-04	0.772296	> 0.01
Ackley	100	1.42E+00	±4.57E+00	1.51E+00	±4.62E+00	-0.0758571	> 0.01

References

1. T. Bartz-Beielstein. Experimental Research in Evolutionary Computation: The New Experimentalism. *Springer*, 2006.
2. M. Črepinšek, S. H. Liu, L. Mernik. A note on teaching-learning-based optimization algorithm. *Information Sciences*, 212:79–93, 2012.
3. J. Demšar. Statistical comparisons of classifiers over multiple data sets. 'newblock *The Journal of Machine Learning Research*, 7:1–30, 2006.
4. J. Miller, J. Daly, M. Wood, M. Roper, A. Brooks. Statistical power and its subcomponents - missing and misunderstood concepts in empirical software engineering research. *Information and Software Technology*, 39:285–295, 1997.

Table 12 Non-parametric Wilcoxon test for different optimization methods (Table 10)[8].

		Čre.[2]	Wag.[8]	Wilcoxon statistics			
Function	D	Mean	Mean	d	$\|d\|$	rank	
Rosenbrock	5	3.55E-02	1.76E-03	0.034	0.034	2	R+ = 12
Rosenbrock	10	9.38E-02	6.92E-02	0.025	0.025	1	R- = 3
Rosenbrock	30	2.16E+01	2.17E+01	-0.1	0.1	3	T = 3
Rosenbrock	50	4.33E+01	4.30E+01	0.3	0.3	5	z = -1.213559752
Rosenbrock	100	9.47E+01	9.46E+01	0.1	0.1	4	
Ackley	5	1.81E-15	9.47E-16	8.63E-16	8.63E-16	2	R+ = 11
Ackley	10	4.23E-15	3.55E-15	6.8E-16	6.8E-16	1	R- = 4
Ackley	30	4.48E-15	3.55E-15	9.3E-16	9.3E-16	3	T = 4
Ackley	50	2.75E-01	4.73E-05	2.75E-01	2.75E-01	5	z=-0.943879807
Ackley	100	1.42E+00	1.51E+00	-9.0E-02	9.0E-02	4	

5. F. Neri, E. Mininno, G. Iacca. Compact Particle Swarm Optimization. *Information Sciences*, 239:96–121, 2013.
6. S-Y. Park, J-J. Lee. An efficient differential evolution using speeded-up k-nearest neighbor estimator. *Soft Computing*, 18:35–49, 2014.
7. D. Sheskin. Handbook of Parametric and Nonparametric Statistical Procedures. *Chapman & Hall, CRC*, 2006.
8. G. Waghmare. Comments on "A Note on Teaching-Learning-Based Optimisation Algorithm". *Information Sciences*, 229:159–169, 2013.
9. F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics*, 1:80–83, 1945.